Accelerating the review of complex intellectual artifacts in crowdsourced innovation challenges

Nagar, Yiftach ; De Boer, Patrick ; Garcia, Ana Cristina Bicharra

Abstract: A critical bottleneck in crowdsourced innovation challenges is the process of reviewing and selecting the best submissions. This bottleneck is especially problematic in settings where submissions are complex intellectual artifacts whose evaluation requires expertise. To help reduce the review load from experts, we offer a computational approach that relies on analyzing sociolinguistic and other characteristics of submission text, as well as activities of the crowd and the submission authors, and scores the submissions. We developed and tested models based on data from contests done in a large citizen-science platform - the Climate CoLab - and find that they are able to accurately predict expert decisions about the submissions, and can lead to substantial reduction of review labor, and acceleration of the review process.

Originally published at:
Nagar, Yiftach; De Boer, Patrick; Garcia, Ana Cristina Bicharra (2016). Accelerating the review of complex intellectual artifacts in crowdsourced innovation challenges. In: Thirty Seventh International Conference on Information Systems, Dublin 2016, 11 December 2016 - 14 December 2016.

Accelerating the Review of Complex Intellectual Artifacts in Crowdsourced Innovation Challenges

Completed Research Paper

Yiftach Nagar
Center for Collective Intelligence
Massachusetts Institute of Technology
ynagar@sloan.mit.edu

Patrick de Boer
Department of Informatics
University of Zurich
pdeboer@ifi.uzh.ch

Ana Cristina Bicharra Garcia
Institute of Computing
Universidade Federal Fluminense
bicharra@ic.uff.br

Abstract

A critical bottleneck in crowdsourced innovation challenges is the process of reviewing and selecting the best submissions. This bottleneck is especially problematic in settings where submissions are complex intellectual artifacts whose evaluation requires expertise. To help reduce the review load from experts, we offer a computational approach that relies on analyzing sociolinguistic and other characteristics of submission text, as well as activities of the crowd and the submission authors, and scores the submissions. We developed and tested models based on data from contests done in a large citizen-science platform - the Climate CoLab - and find that they are able to accurately predict expert decisions about the submissions, and can lead to substantial reduction of review labor, and acceleration of the review process.

Keywords: Open-innovation, Crowdsourcing, Collective-Intelligence, Filtering, Evaluation, Sociolinguistics, Social-computation

Introduction

On April 20th 2010, around 10pm local time, the DeepWater Horizon oil rig exploded in the gulf of Mexico. Oil started leaking from the well it was drilling, and this leak soon developed into the largest oil spill to date. Recovery efforts involved over 47000 people from 19 countries, and yet, resources alone were not enough. Dealing with the sheer magnitude of the leak, and with the unprecedented complexity of sealing the well deep below the sea required ingenuity. The general public was tapped for ideas, and responded: Before long, the organizers of the recovery operation received over 120,000 submissions!

In order to evaluate them, and to select the best ideas for implementation, a special assessment team counting more than 100 experts was established. It took them several months to review all submissions. But could they – or other organizers of crowdsourced ideation and innovation challenges – filter submissions more efficiently in order to find the most promising ideas faster, and with less effort?

In this paper we present an approach that may help.

Crowdsourced Innovation Challenges

The crowdsourced innovation effort during the oil spill is but one example of a growing phenomenon. While many problems are routinely solved by individuals or teams within the boundaries of traditional organizations, the knowledge and skills needed for solving some very complex and/or novel problems are distributed widely outside the organization, sometimes in places that are not known in advance (Chesbrough 2003; Malone et al. 2009). Seeking ideas and knowledge widely outside of the organization can help organizations discover radical new ways of thinking about their problems, and may lead to creative solutions (Lakhani et al. 2007; Terwiesch and Xu 2008). Indeed, people coming "from the outside", who bring different perspectives and heuristics, often generate the best solutions to innovation challenges (Jeppesen and Lakhani 2010). These realizations, and growing evidence of the value of open innovation, led many companies, governments, and non-profits to adopt an open innovation approach.

As part of their open innovation strategy, organizations increasingly develop and run innovation contests (cf. Boudreau et al. 2011; Boudreau and Lakhani 2013; Morgan and Wang 2010; Terwiesch and Xu 2008), in which they elicit intellectual artifacts including ideas, plans, designs and predictions from both experts and lay crowds, in search of solutions to various large-scale challenges. In many cases, these efforts yielded solutions to problems where previous attempts have failed (Lakhani et al. 2007; Marshall and magazie 2012). Among other notable examples of successful contests, the Ansari X-Prize, which was won by a team that developed a spaceship able to carry three people to 100 kilometers above the earth's surface twice within two weeks (cf. Haller et al. 2011), and the Netflix prize challenge (Bennett and Lanning 2007) received considerable attention.

In this paper, we primarily focus on crowdsourced innovation challenges and contests, i.e. open innovation challenges that encourage broad participation, and are conducted under competitive terms. While in most cases these are framed as contests (as in the case of the X-prize, the Netflix prize, or in the case of many challenges posted on such platforms as InnoCentive's), there doesn't necessarily have to be a prize involved. What we mean by competitive terms is that the "seeker" – the organizations who initiated the challenge – has limited resources; whether these resources are prize money, or, as in the case of the oil spill, resources (including time, people, tools, money, managerial capacity, etc.) to pursue and implement the most appealing offered solutions.

The designers of these challenges and platforms overcame a critical challenge: to motivate and harness intellectual work of large crowds of people (in some cases many thousands), using incentives that appeal to intrinsic and extrinsic motivations. Some designers also invented novel ways of organizing the work of crowds of people such that inputs from many (often: lay) people can be validated, refined, and combined, to yield outcomes of surprisingly high quality. However, receiving a lot of submissions from a large and diverse crowd poses a new challenge to designers and operators of crowdsourced innovation contests: evaluating a mass of complex intellectual artifacts, and selecting the best among them (Nagar 2013).

The bottleneck of expertise in the evaluation process

In some crowdsourcing systems, evaluating crowd inputs is straightforward. For example, reCAPCHA (Von Ahn et al. 2008) uses simple statistical calculations to processes many millions of human attempts of recognizing obscured scanned words each day. In FoldIt (Cooper et al. 2010), a system that asks the crowd to assists in predicting and designing protein-folding structures, domain expertise was coded into the system such that each proposed way of folding a protein can be computationally evaluated in an instant.

However, evaluating intellectual artifacts becomes a significant challenge when those artifacts are complex (e.g. when they contain a lot of unstructured text), when no computational methods are applicable, where domain expertise is scarce, or where criteria for evaluation are themselves complex and ambiguous (what is "novel" for example, is given to interpretation). For instance, the assessment of proposed solutions to scientific challenges posted by InnoCentive requires high levels of expertise that is not readily codifiable, and which is only available in the heads of domain experts; a resource that is both scarce, and expensive. Or, consider the review of grant applications by the US National Science Foundation (NSF): according to Boudreau et al. (2012), in 2010 alone the NSF brought over 19,000 scientists to the Washington DC area to participate in proposal evaluation. Beyond the potential to incur significant costs, the bottleneck of expertise for performing evaluation and selection can cause substantial

delay in finding solutions. In September 2008, a company named Google launched a crowd-innovation challenge called "Project 10^{100}" that asked people to submit ideas that have the potential to change the world. According to Google's original plan, winners were supposed to be announced in January 2009, following a 3-month review cycle. Yet, after receiving over 150,000 submissions, Google had to postpone the announcement of winners multiple times, and eventually, announced them only in September 2010, twenty months(!) after the original planned date. During that time, thousands of Google employees took part in the review process. As these examples clearly illustrate, the bottleneck of expertise for performing evaluation and selection can become one of the most critical hurdles for collective-intelligence systems addressing large problems.

Relieving the Bottleneck of Expertise

Relieving this bottleneck of expertise in the review process is not of the type of challenge that can simply be "solved". But we believe that research and innovation in ways of organizing review work, as well as in complementary computational approaches, such as the one we introduce in this paper, can lead to significant improvement over the current state of the art. Before discussing our approach, we review related work.

Related Work

The current state-of-the-art in reviewing submissions is not much different from the state-of-the-art a half-century ago: in the field, the cumbersome, labor-intensive, slow process of expert panels, review committees, and variations thereof, is still dominant.

More recently, some attempts were made to relieve the bottleneck of expertise for reviewing; but to our knowledge no new method has been widely adopted yet. These attempts generally fall into two categories: organizational (mainly using crowdsourcing), and computational.

Crowdsourcing evaluations

Crowd-based evaluation and filtering systems come in several variations: crowds (which might include organization members, volunteers, or paid crowds in online labor markets) are recruited and are asked to *vote* (e.g. Bao et al. 2011), *rate* (e.g. Blohm et al. 2011; Riedl et al. 2013; Salminen and Harmaakorpi 2012), or *rank* submissions (e.g. Salganik and Levy 2012), based on one, or multiple criteria (Dean et al. 2006). *Prediction markets*, in which crowd members trade contracts based on their beliefs about the likelihood of ideas to be successful, have also been proposed as a way to incentivize and aggregate ratings from the crowd for predicting the quality or success of ideas (Bothos et al. 2012; Soukhoroukova et al. 2012).

Although large crowds may be relatively easily recruited to off-load experts, these methods have all been shown (both theoretically and empirically) to have flaws and limitations (see Klein and Garcia 2013 for a more nuanced and elaborate review). The main limitation (though not the only one) common to all of these approaches is that the size and complexity of the task of comparing alternatives rises rapidly with the number of options that need to be compared. This renders them problematic for use in practice in crowd-innovation challenges where the number of ideas is large, especially when the proposals are complex.

Computational evaluations

Automatic Essay Scoring algorithms (also known as Auto-graders) have been in use for some time now, and recently received renewed attention with advances in natural language processing, and the growing need for such tools for Massive Open Online Courses (MOOCs) (Markoff 2013). Indeed, some positive results have been reported (Shermis and Hamner 2013). However, these tools usually require to be trained on large corpora of manually annotated student essays, which, in turn, are assumed to be somewhat similar to one another. By the very nature of innovation challenges, submissions are very diverse, and creating annotated sets is a labor intensive, slow process. It is also not clear whether and to what extent the rules developed in some setting will be applicable in other settings (Klein and Garcia 2013). For these reasons, automatic essay scoring tools have not been applied to judging submissions in

open innovation contests. Another approach, closer to the one we present in this work, is to develop metrics of the quality of submissions based on word frequency statistics. For example, Walter and Back (2013) measured the use of unique sets of words in order to detect innovative ideas, with mixed results. Although we find their approach promising, the average submission in their setting had 25 words only. Therefore, implications from their study for systems in which submissions are more complicated, may be limited. Westerski et al. (2013) developed an elaborate domain-independent taxonomy for idea annotation, and offer that idea originality and idea dependability (the level to which it is connected to other ideas), are strong predictors.

These results are encouraging, and yet, we believe that there is room for additional work. We present our method in the following section.

Method

We propose a novel computational approach that can help reduce the cognitive load of expert judges, by performing initial screening and/or prioritization of the review queue. Our approach is unique in that it is the first, to our knowledge, to model both the submitted artifact itself and traces of human activity relating to the submission; and in considering how sociolinguistic aspects of the submission text may influence expert reviewers. We tested our approach in context of the Climate CoLab – an open-innovation citizen-science system, and we show it can yield significant improvement in the process.

Approach

Our approach was shaped by several realizations regarding different aspects of the problem:

1. Predicting the winners of open-innovation contests is hard and tricky even for experts. An easier approach is to differentiate high quality from low quality submissions. Such a "triage" step is performed manually in many systems, and does not really necessitate high-level expertise to perform. It is probably the place where computational means can achieve the most reliability, and most impact, by filtering out low-quality submissions, and freeing the experts to devote their time and skill to consider more promising submissions.
2. Although the content of submissions can be computationally modeled in various ways, true understanding of complex ideas is still the realm of humans. However, form matters too, and can be reliably assessed computationally, based on solid theoretical foundations.
3. While expert-panels are notoriously prone to many types of bias, due to a lack of a better alternative their judgment is still de-facto the state of the art, and widely accepted as the gold standard in studies.
4. Previous attempts to computationally classify and rate crowd-proposals relied solely on proposals' text (e.g. Walter and Back 2013; Westerski et al. 2013). Yet, specifically in open-innovation environments, traces of crowd and author activities in relation to the proposals are available, and may provide additional clues that can help predict which proposals would be favored by expert judges.

Our resulting approach is open-ended. Based on data available in our setting, we devised a preliminary taxonomy of variables which can serve as a guideline for modeling work, and which can be enhanced and appropriated to fit different settings. We developed and tested models based on this taxonomy, which take into account sociolinguistic and other aspects of proposals' text, as well as author and crowd behavior. With these models we aim to match the reviewers decisions at the first triage stage. We demonstrate our approach in the context of one platform – the *Climate CoLab*.

Setting: the Climate CoLab

The Climate CoLab[1] (Introne et al. 2011) is a sociotechnical system designed to help thousands of people around the world collectively develop plans for addressing global climate change. The CoLab combines several design elements, including model-based planning and simulation, and a crowdsourcing platform

[1] http://www.climatecolab.org

where citizens work with experts and each other to create, analyze, and select detailed proposals for what to do about climate change. As of September 2016, over 200,000 people have visited the Climate CoLab, representing virtually every country in the world, and over 70,000 have registered as members. The main activity under the CoLab is a set of ideation contests on a range of topics, from how to reduce emissions from electric power generation to how cities can adapt to climate change. Past winning proposals have been presented to decision makers in the UN and the US congress, and to potential implementers. In the 2012-2013 set of contests that we analyzed, beyond the announcement of winners in each contest, a grand a prize of USD 10,000 was granted to the proposal that was selected best across all contests.

The current review process at the CoLab

Proposals are submitted on the CoLab's website, using a template which asks authors to indicate the what, where, who and when of the proposal. Proposals can include text, as well as multimedia, and while some proposals are incomplete, or of low quality - many high-level proposals are submitted as well. To select the best ideas, the CoLab staff has developed an ad-hoc review organization that includes volunteers in two roles: *Fellows* are graduate students and young professionals; and *Expert Judges* – mainly senior faculty, and industry veterans. Fellows and judges are recruited for specific contests, based on their expertise in the contest topic. After the fellows perform initial screening, they, together with the expert judges, select "Semi-finalists". Authors of semi-finalist proposals are given a chance to revise their proposals, and after that, judges and fellows select the finalists, from which winners are later selected. This process is very labor-intensive. For reviewing the set of 2012-2013 contests, about 60 volunteer reviewers (about half of them *fellows*) were recruited.

In addition, the crowd (i.e. the CoLab community of registered users) can make comments on proposals and indicate their support for a proposal by clicking a "thumbs-up" button (akin to the "like" action on online social-networking platforms). During the last phase, the crowd also votes to select the crowd-vote awards.

The datasets

We analyzed two datasets, comprised of the proposals submitted to contests that ran under the Climate CoLab framework in 2012-2013, and in 2013-2014. These contests covered a wide range of both technical and social topics related to dealing with climate change, such as the reduction of greenhouse gas emissions from transportation systems, geoengineering to avoid methane feedback, and urban adaptation. A complete list of the contests and of the proposals is available on the Climate CoLab website. Key descriptive statistics of these datasets are presented in Table 1. Of these proposals, about 22% were selected as "Semi Finalists" by the CoLab's fellows and judges (81 in 2012-13, and 116 in 2013-14). These semi-finalists were reviewed more thoroughly, and their authors received detailed constructive comments, and were given an opportunity to submit revised versions. In each year, following another revision cycle, the finalists were selected from the semi-finalist pool, and eventually, winners were elected.

We focused attention on the *first* stage of the review, i.e. the selection of semi-finalists.

Our analysis was not done all at once. We first analyzed the 2012-2013 dataset when it became available, and later, the 2013-2014 dataset when it became available. As explained below, we introduced slight enhancements in the later analysis.

Table 1. Datasets

Contest year	**2012-13**	**2013-14**
Number of contests	18	18
Number of proposals analyzed	369	510
Number of 'Semi-Finalists'	81	116
Number of Expert Judges; Semi-expert 'fellows'	30; 30	60; 30

The metrics

These include the data of the proposal itself, and activities of authors, community members and the crowd in relation to the proposal. Naturally, some metrics ended up as more useful, i.e. providing stronger predictive power, while some metrics we tested were not as predictive. But rather than detail only the most predictive metrics that ended up in our final model, we chose to review all metrics we looked at, because a part of the contribution of this paper is to provide others with a broad taxonomy of metrics that can be observed in different systems, and which may prove useful in other settings. We grouped the metrics into six categories as detailed below.

Readability of the proposal text

Numerous studies show that easier reading improves comprehension, retention, reading speed and readers' perseverance (also called depth or persistence: the tendency to keep reading the text) (DuBay 2007). The ease with which a reader reads a text depends on the reader's skill, knowledge, interest and motivation, as well as on features of the text: its content, style, design and organization (DuBay 2007). Early in the 20th century, *educators started using vocabulary difficulty and sentence length to predict the difficulty of a text. This has spurred intensive research and development of readability formulae.* These formulae use counts of language variables in a piece of writing in order to provide an index of probable difficulty for readers (Klare 1974). Many readability formulae have been developed over the years, and while there has been critique about their misuse and their value in certain applications, DuBay (2007) notes that they *"have proven their worth in over 80 years of research and application"*.

Our first hypothesis therefore is that readability may influence human expert judges as they read proposals, and specifically, that low readability will hinder the proposal's chance of being favored by the judges.

We did not define as a goal to find the "best" readability index that would provide the highest correlation with proposals success. There are literally hundreds of readability indices, but they are better thought of as rough guides than as highly accurate values (DuBay 2007). To check whether our intuition about readability has merit, we selected four indices that are in very common use in many applications: Flesch-Kincaid Grade Level, Flesch Reading Ease (Kincaid et al. 1975), Automatic Readability Index (ARI) (Kincaid et al. 1975), and the Coleman-Liau Index (CLI) (Coleman and Liau 1975).

Writing style: Function words and language style matching

Style words, also known as *function words*, including pronouns (such as *I, you, they*), articles (*a, an, the*), prepositions (*to, of, for*), auxiliary verbs (*is, am, have*), and some other common word categories, account for more than half of the words that occur in human communication (whether written or spoken) (Pennebaker 2011). While these words convey very little meaning on their own, extensive research demonstrated that by analyzing their use, we can learn about the personality, social skills, honesty and intentions of the people who use them (Pennebaker 2011).

Further, social psychologists and sociolinguists, who study the use of language in social contexts, have shown that people match their language, stylistically, to that of other people with whom they are communicating. Researchers have further shown that a reliable index of language style matching can be constructed by using counts of function words. This index also correlates with social-psychological phenomena such as the strength of dyadic relationships, group cohesiveness and group task performance (Gonzales et al. 2009; Ireland and Pennebaker 2010).

It seems plausible that language style might also affect expert reviewers' perception of the proposals, and influence their decisions. While we did not have writing examples from the reviewers that may have allowed us to check the matching between their writing styles and those of proposal authors, the pool of reviewers of the Climate CoLab can be characterized with some common traits: highly educated, highly conscientious, working in academia or in knowledge work. We conjectured therefore, that as a collective, CoLab reviewers tend to have similar stylistic preferences regarding the writing of the proposals. For instance, we hypothesized that they will prefer to see proposals that are written in more 'academic', rather than colloquial style.

It is not easy to directly map such a hypothesis to specific function words a-priori. We took an exploratory approach, and – rather than selecting a subset of words based on any theoretical basis – let the data speak. We used the 2007 version of LIWC (Pennebaker, Booth, & Francis, 2007) – a linguistic analysis tool, which calculates the percentage of words in a text that fall into each of 15 function word categories (several of which overlap hierarchically, e.g., first-person singular pronouns are a subcategory of personal pronouns. See Pennebaker et al. (2007) for a complete list of variables and further details).

We focused on variables whose values in the semi-finalists group were statistically-significantly different from their values in in the non-semi-finalists set (based on Mann-Whitney's two-tailed test, $\alpha = 0.05$). We thus narrowed down the list to 17 variables[2], which are depicted in Table 2.

Table 2. Remaining LIWC variables

Category	Variable	Examples
Cognitive Processes	Discrepancy	should, would, could
	Inclusive	And, with, include
	Negate	No, But
Linguistic Processes	Auxiliary verbs	Am, will, have
	Common verbs	Walk, went, see
	Dictionary words	
	Personal pronouns	I, them, her
	Present tense	Is, does, hear
	Total function words	
	Total pronouns	I, them, itself
	Words>6 letters	
Personal concerns	Achievement	Earn, hero, win
	Money	Audit, cash, owe
	Work	Job, majors, xerox
Punctuation	Commas	,
	Dashes	-, –
	Parenth	(,)

Potential indicators of the completeness and maturity of the proposal

We have described above the mixed blessing of asking the crowd to submit ideas. On the one hand, with the right incentives, many more ideas are submitted than would have otherwise, raising the likelihood of finding diamonds in the rough. On the other, since the crowd is diverse, and includes people with different levels of relevant knowledge, skill, and motivation, the quality of submissions varies greatly, and the quality of many submissions may be poor.

We hypothesized that several metrics of the text might signal how much work was put into creating the proposals, and accordingly, the completeness and maturity of the proposal:

1. The number of references (NumReferences): the last section in the CoLab proposal template allows the authors to include references to external sources. We hypothesized that more references can signal that more work was done on the proposal. (The reference lists on CoLab proposals are much shorter than those in academic papers, where such a relation is less likely to hold. While the maximum is 49, the mean number of references in our corpus is 4.11 and the median is 1. About 40% of the proposals had no references in the reference section.

[2] Although word count is included in LIWC, in our modeling we assigned it in a separate category (proposal length, see below).

2. The number of hyperlinks (NumHyperlinks)
3. The number of images
4. Whether some sections were left unfilled. The proposal submission interface does not force authors to fill all the sections of the proposal. This is done deliberately, to allow people to submit "half-cooked" ideas, and allow others to respond and assist. As a result, some proposals remain in this state when judging starts. This would not necessarily disqualify them, and it is possible (though less likely) that a proposal can advance to the semi-finalist stage even if not all its sections are complete. We built a set of dummy variables to indicate whether any section was empty.

Length

Length affects readability, it is related also to style, and can potentially indicate something about proposal maturity. Since length is related to all 3 categories above, we decided to treat it as a separate category. It seems likely that immature proposals, e.g. proposals that have complete sections unfilled, would be shorter. We therefore assumed that very short proposals would have a lower chance of being selected. It is not so clear, however, that the relationship is monotonic. One could assume that very long proposals might be frowned upon, at least by some judges.

Length can be measure by the number of letters, words, sentences and paragraphs. We measured all of them.

Crowd activity

In the early stages of the contest, members of the community can indicate their support for a proposal by clicking a "thumbs-up" button (akin to the "like" action on online social-networking platforms). In addition, members of the community (including fellows and judges), as well as authors, and everyone else, can comment on proposals during all phases of the contest.

We initially considered the following metrics when analyzing the first dataset:

1. Number of comments: on the one hand, it seems likely that more interesting ideas will receive more attention, and drive more engagement, which will be positively correlated with comments. On the other hand, it seems likely that people comment when they see flaws, more so than they do just to say words of support. It was therefore not clear to us whether we will see a strong correlation with the outcome, yet we thought this was worth checking.
2. Number of comments made by experts: we checked whether comments made by members of the review team at an early state before the deadline were correlated with later selection of proposals as semi-finalists.
3. The proportion of comments made by experts.
4. The number of "Likes" the proposal received from the community: The CoLab community is an unusually highly-educated community, and yet, on average, members do not have the same level of expertise as the judges. Although it is likely that expert judges will judge proposals somewhat differently from the average member of the community, we still expect to see some correlation between the "taste" of the community, and that of the judges.
5. Proportion of "Likes": Because our data set includes proposals that were submitted to 18 different contest, we adjusted the number of likes, to control for the differences in the number of proposals across contests, and the number of people interested in them. Some contests drew more proposals and more crowd activity than other contests. The proportion of likes is the number of likes a proposal received, divided by the total number of likes received by all proposals in that contest.

For the 2013-2014 contest cycle we were able to receive data from the web-analytics software installed on the Climate CoLab web servers. Thus, we added the following "honest signals" of crowd activity to our taxonomy:

6. Pageviews (indicates how many times the web page with a proposal was viewed);
7. uniquePageviews (indicates the number of unique IP addresses from which a web page with a proposal was viewed);
8. AverageVisitTime (the average time each proposal page was viewed).

9. LikesPerView (the number of likes the proposals received, divided by the number of views it received);
10. LikesPerUniqueView (the number of likes the proposals received, divided by the number of unique views it received);
11. VisitTimeVsWords (the average time spent on the proposal page, divided by the number of words in the proposal).

Author actions

1. Number of Days: the number of days left between the initial submission and the submission deadline. We have heard speculation and occasional observations from organizers of the Climate CoLab as well as other prize-bearing crowdsourcing ideation challenges, about strategic behavior of some authors, who submit their proposals close to the deadline, seemingly to prevent others from copying their ideas. One contest organizer conjectured that the best proposals are among the last to be submitted, though he had no supporting data. Although it is indeed the case that most proposals are submitted very close to the deadline, this could be the result of mere procrastination. We therefore decided to check whether there is a relationship.
2. Number of Updates: Once a proposal is submitted, authors can update it (as well as let other members of the crowd to do so) as many times as they want. Can the number of updates help predict which proposal will be selected as a semifinalist? More updates may mean that the proposal was not well thought of in advance, but they could also mean that more work is being done.

Modeling

Our statistical analyses began with non-parametric correlation tests[3] that helped us identify the variables that would be good candidates for our models, as well as to avoid issues of multicollinearity in our models. We then created a series of logistic-regression models for each category of variables, using partial sets of variables that were not strongly correlated with each other, and came up with the best model[4] for each category after eliminating variables that did not have statistically-significant effects on the outcome variable. We then constructed a set of integrated models, which combined the most salient predictors from all categories, and selected the final model. Finally, we validated our model by building a machine-learning classifier and performing a stratified 10-fold cross-validation.

Results

Logistic regression results

In Table 3 we depict a partial summary of our modeling for 2012-13, comparing the final models in each category, and the final integrated model. This illustrates our modeling process, and shows how the integrated model is (as expected) better. In Table 4 we present a partial summary of our modeling for 2013-14, showing several integrated models.

[3] We used the Kendall-tau correlation test, since not all of the variables are normally distributed in the dataset.

[4] We compared the models based on measures of goodness-of-fit and predictive power, including -2*Log-Likelihood (-2LL), Akaike Information Criterion (AIC), Pearson Chi-square, and the Area Under the ROC Curve (AUC).

Table 3. Representative logistic regression models from different categories, and final integrated model (2012-13 data)

	Readability	Writing Style	Maturity	Length	Crowd Activity	Author Activity	**Integrated**
(Intercept)	-2.694***	-0.305	-2.641***	-2.534***	-0.984***	-0.984***	**-2.964*** **
CLI	0.091**						
ppron		-0.229*					**-0.379** **
verb		-0.081*					
References			0.057**				
Timeline			1.389***				**1.284** **
NumWords				0.001***			**0.0001*** **
PropLikes					4.129***		**3.517** **
NumDays						-0.007**	
-2LL	376.91	374.42	356.41	339.39	376.75	376.12	**313.04**
AIC	380.91	380.42	362.41	343.39	380.75	380.12	**323.04**
p>χ^2	0.0007	0.0009	1.13E-07	2.55E-12	0.0006	0.0005	**1.67E-15**
AUC	0.658	0.647	0.733	0.796	0.680	0.637	**0.821**

Table 4. Representative integrated models and final Model (2013-14 data)

	C1	C2	C3	**C4**	C42	C43	C5
(Intercept)	-1.706	-2.044**	-1.393***	**-**	-1.372***	-2.077***	-
FleschKincaidGradeProposal	0.029	0.058					
SmogProposal	0.038						
GunningFogProposal	0.002						
Parenth	0.392*	0.369*	0.401*	**0.402***	0.572***		
Negate	-1.418***	-1.477***	-1.569***	**-**	-1.452***		
Money	0.164**	0.173**	0.167**	**0.155***	0.181**		
TotalSectionsNotEmpty	-0.052						
NonEmptySections<=6	-1.684*	-1.498**	-1.574***	**-**	-	-1.636***	
NumLettersProposal	0.*	0.~	0.*	**0.** **		0.***	
PropLikes	4.914**	5.436**	5.5**	**5.638*** **	6.622***	5.674***	6.569***
PropExpertComments	-0.406	-0.376					
LikesPerUniqueView							
VisitTimeVsWords	-1.518	-1.566	-1.726				-4.737***
-2LL	421.332	427.786	431.168	**433.885**	444.247	464.253	493.562
AIC	447.332	447.786	447.168	**447.885**	456.247	472.253	563.562
p>χ^2	2.91E-19	2.03E-21	5.92E-22	**4.76E-**	1.46E-20	8.30E-18	2.62E-12
AUC	0.8134	0.8124	0.8089	**0.8085**	0.7916	0.7737	0.7519

Interpretation of the final models

Length proved to be a very strong predictor. In both contest cycles (which for the most part included different contests and different judges), longer proposals had a better chance to be selected as semi-finalists (and it is surprising how close the weights of this parameter were: an increase of 10,000 characters was equal to an increase of {~10.6%; 12.6%} for a proposal to be elected semi-finalist in our {2012-13 model; 2013-14 modes}, ceteris paribus.

Proposal completeness is important. Although the CoLab contests yield proposals that experts have deemed as very high quality, many submissions are immature, and even incomplete. E.g. about 30% of the submissions in the 2012-13 cycle have left the "timeline" section in their proposals empty, though it is part of the submission template. The fitted odds-ratio of a proposal which included information in the Timeline section to be selected (compared to a proposal with no timeline) was 3.61. Of 81 semifinalists, only 8 were proposals that had no timeline. In the 2013-14 analysis we extended our analysis to observe other missing sections and found that omitting more than half the sections was a very strong indicator that a proposal would not be a semi-finalist.

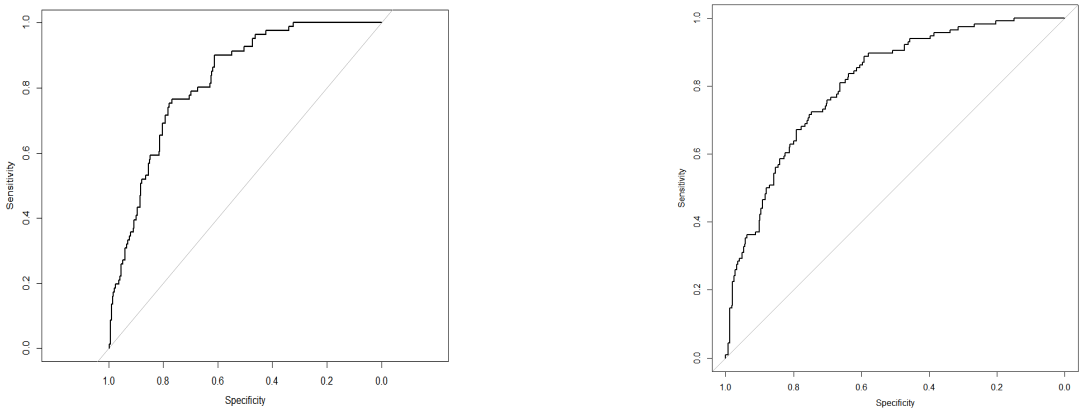
A higher proportion of "likes" also indicated higher likelihood for the proposal to be selected, though we should be careful with stating an odds-ratio as this estimate had a large confidence interval.

Linguistics: some linguistic features were predictive (e.g. pronouns were negatively correlated with success in 2013; the usage of words related to money was positively correlated with success in 2014). However because these effects did not carry over the two datasets we should be careful with stating strong conclusions about them.

Evaluation of model performance, and implications for implementation in practice in the CoLab and in other settings

To validate and evaluate our model, we built a Logistic Regression classifier with scikit-learn (Pedregosa et al. 2011) using the variables we have selected in our combined model, and ran a stratified 10-fold cross-validation (as advised in Kohavi 1995). The resulting model accuracy for 2012-13 was 0.789 (±0.06) and the average AUC was 0.816 – very close to the AUC of the combined model in Table 3. ROC Curves[5] of the selected models for both years are shown in Figure 1.

Figure 1. ROC Curves for final models: Left (2012-13), Right 2013-14



To illustrate the strength, and potential implications of our results, let us consider a couple of examples, by selecting cutoff points on the ROC curve. The selection of such points depends on the preferences of the system designers, i.e. the organizers of each innovation challenge. The nature of many of these challenges is such that the preference of their organizers would be not to miss a good idea that might be

[5] ROC stands for Receiver-Operator-Characteristic (Swets 1988; Swets and Pickett 1982; Zweig and Campbell 1993)

"hiding in the haystack". This is often the main motivation for organizing such open innovation challenges in the first place, and indeed this is the case with the Climate CoLab. The cost of reviewing proposals is marginal compared to the benefits we can reap from solutions to problems caused by climate change. That said, operation costs and resources are not unlimited, and the time of expert reviewers, whether they volunteer or they are paid, is a scarce resource, hence the need to reduce the demand for expert time.

Assume, therefore, the case where no good idea should be missed. We will then tune our threshold to maximize sensitivity, on account of specificity. For the purpose of this illustration let us use numbers from the 2012-13 final model. We can select our cutoff rating to be p^=0.033. Results from the model are depicted in Table 5.

Table 5. Model output, threshold = 0.033

	Selected	Not Selected
Indicated by the model	81	236
Not indicated by the model	0	52

The sensitivity of the model in this case is 100% (all semifinalist are indicated by the model). But the model is also able to correctly identify 52 non-finalists. This means that even under the strictest circumstances, where the model is tuned to 100% sensitivity, it could have reduced the amount of work of experts by filtering out 52 lower-quality submissions (about 14% of all submissions). In a system like the Climate CoLab, which employed over 60 volunteering experts and semi-experts who spent many days reviewing, such a relief can be substantial.

Additional gains can be made by tuning the model to increase specificity. Assume, for example, another scenario, where the cutoff point is set on p^=0.185. Results from the model in this case are depicted in Table 6.

Table 6. Model output, threshold = 0.185

	Selected	Not Selected
Indicated by the model	71	112
Not indicted by the model	10	176

In this case, the model sensitivity drops to 87.7%, but specificity goes up substantially: the model would indicate 186 submissions as submissions that are not likely to be selected by expert human judges to be semi-finalists, and would be correct about 176 of those. What about the 10 semi-finalists that we will incorrectly classify as lower quality? When considering the use of a model and cutoff points, we cannot guarantee either 100% sensitivity or specificity. But implementing a prediction model such as our model can allow contest organizers to organize the review process somewhat differently, and hopefully, more efficiently. For instance, instead of assigning the entire proposal pull to be reviewed by expert judges, they can allocate the review of submissions that were highly-rated by the model (and therefore believed to be more likely to be deemed as higher quality by the judges) to experts, and the review of proposals that were low-rated by the model, to lesser-expert reviewers. In such case, the size of the highly-skilled expert panel can be dramatically reduced (e.g. in the case of the second scenario, the initial pool of submissions sent to high-level experts in this case would have been only 71+112=183. That's a reduction of about 42%, from the 317 proposals we would send to the experts if we choose the more conservative threshold, and of about 50% from the baseline of not using a model at all. Alternatively, contest organizers may keep the review process in the hands of experts, but prioritize the review sequence such that submissions which received lower scores by the model would be reviewed later. This approach can be helpful in cases where the crowd's help is asked in response to crisis, and speed is of essence, such as the crowdsourcing effort that followed the "Deep Horizon" disaster in the gulf of Mexico in 2010.

Generalizing to new settings

The strong predictive power of the models we offer is obvious. However, it is a legitimate question, of course, what we can take from this to use in other settings. The answer has several parts: First, the taxonomy of potential predictors we offer here is not unique to the specific setting we studied. Readability measures, linguistic analysis, checks for completion and maturity and length are applicable to any textual submission. Author and crowd activities are not always recorded and not always relevant, but might be in some settings. We offer this taxonomy as a basis for at least initial model building. Of course setting-specific variables can be added in context.

Second, although we analyzed data from one platform, the data came from 36 different contests, judged by dozens of judges, and we see that at least some of the variables seem robust. Another robustness check we did was to run the integrated model we built for the 2012-13 dataset on the 2013-14 dataset. The model performed well, yielding an AUC = 0.75 when rebuilt on 2013-14 data, and – perhaps even more impressively – yielding an AUC = 0.74 when used with the parameter weights that we derived on the 2012-13 dataset. This is not too far from the best 2013-14 model performance.

Third, and importantly, we performed a bootstrapping analysis, to check whether our approach can be used to build a powerful model by using only a subset of previously-judged proposals. The results, shown in Table 7, are encouraging.

Table 7. Bootstrapping results

% of data used to build the model	Resulting Area under ROC curve
80%	0.81
40%	0.80
20%	0.73
5%	0.72

Thus, while we do not assume that the specific weights from our modeling will generalize to all settings, and further – perhaps some variables will change in the models, the results of our bootstrapping analysis suggest a path for implementation in new settings, which does not mandate waiting for one cycle of contests to end: a small set of rated proposals can be used to build a solid preliminary model with good performance that can already save work. Small "batches" of proposals rated highly by the model can then be judged by expert judges, and small batches of proposals which the model rated low, can be inspected by people with lesser-expertise just to make sure good submissions are not prematurely dismissed without proper review. Feedback from these human judges can then be used to fine-tune the model, whether by carefully analyzing the reasons for model errors, or automatically, by implementing a learning scheme into the model (e.g. by building a Bayesian classifier). We intend to further examine this approach in our future work.

Lastly, it is clear that further experimentation/analyses with data from other settings will be helpful. We are working towards that end, and are open to collaborate.

Discussion

We are proposing a mechanical approach that can assist human-experts by automatically scoring these complex intellectual artifacts. Experts may be able to use these scores as indicators that can assist them in screening the initial pool of submissions, freeing them to dedicate their time to consider the better submissions. Our open-ended approach, which relies on integrating various features of the artifact, with various pieces of data from human activity relating to the proposal has yielded very promising results when applied to data from the Climate CoLab.

Most of the specific results from our models would not, by and large, be a huge surprise to people who have acquired some experience in running crowd-innovation challenges. They re-affirm some tacit

knowledge, e.g. that the quality of a lot of submissions is low; that more complete, more mature proposals, written in a more formal way have a higher chance of being favored by experts. The contribution of this paper lies not in highlighting these relations, but rather in:

1. Suggesting an open-ended approach that makes use of data from multiple sources, including the artifact and human activity of the crowd;
2. Offering an initial taxonomy that can serve to guide people interested in building predictive models for additional settings, that can lead to significant, tangible improvement in the review process; Specifying many variables in this taxonomy and demonstrating how they can be measured;
3. Empirically demonstrating the implementation of the approach, and its predictive power, using field data from a real live platform.

Additional consideration for practice

The approach we propose should be carefully appropriated, tuned, and tested to fit different settings. Not all variables we have examined may be available or easy to obtain in different settings. Yet, even simplistic and minimal use of some of the measures can be useful. For example, we noted that word count alone would be a very strong predictor in our setting. By setting a threshold of 100 words, we could eliminate 38 submissions without throwing away any submission that made it to the semifinals. That is already 10% of the submissions. Raising the bar to 250 words would have eliminated 71 proposals (19%), with only one proposals that was later selected as a semi-finalist, and a threshold of 500 words would have indicated 120 proposals (32.5% of the entire pool) as low-quality (with 3 false positives).

One potential reason to prefer the inclusion of additional variables in the model, rather than strictly preferring the most parsimonious model, is that some of these additional variables can help address a concern we have heard from some scholars and contest organizers with whom we have discussed our work, regarding fraud-attempts. "If you publicize your model", they asked, "wouldn't it cause some people to try to game the systems (e.g. by writing longer proposals)"?

Because our approach relies on multiple criteria, and since at least some of them are difficult to manipulate, e.g. crowd behavior, and writing style (Ireland and Pennebaker 2010), our approach is quite resilient to attempts to game the algorithm. It is further worth noting that since we propose to use our method for filtering the lower-quality proposal, and leave the evaluation of good submissions to experts, any attempt to artificially game the system is economically senseless, since if any such bogus submission would deceive the algorithm, it would eventually be reviewed by an expert, and dismissed.

Another concern that we heard from some scholars and practitioners was whether the community of solvers would accept usage of the kind of system we propose here as legitimate, and whether it would not hurt the motivation of potential participants and deter people from participating. This is a fair question.

Our answer goes back, again, to the way we propose to use the algorithm: not to replace the experts, but to support them. We believe that in an environment of complex, nuanced solutions to complex, nuanced, multi-faceted problems, human experts should be the ultimate judges among the best proposals. The algorithm we developed does not aim to understand submissions and judge them in a meaningful way, but rather – to approximate crude judge decisions (high-quality vs. low-quality). The option to override can always stay there (if contest organizers have the resources).

It is in the hands of contest organizers to decide what level of filtering they would like to operate, based on the human expert resources available to them. We have suggested above one modus operandi in which semi-experts (e.g. students) can cross-check submissions that the algorithm marked as low quality. Another way can be to set an appeal mechanism. The algorithm is not a standalone tool, but rather, a tool that should be used in context. If used this way, and if the reasons for why it is used, and the way it is used are communicated candidly and transparently to the community, we believe that it should be accepted. Going back to our opening example of the oil spill: we believe that if such an algorithm would have been available and used at the time, and if it would have been made clear to the public that it is used to help the experts *prioritize* their review work, in order to raise the chances of minimizing the environmental damage from the leak, this would have been widely accepted as a legitimate approach.

Conclusion

The bottleneck of expertise for reviewing a mass of complex ideas submitted to crowdsourced innovation challenges is a real and painful problem. As this model of open-innovation – which has already proven useful in finding solutions to hard problems where none existed before – becomes more prevalent, the need to relieve this bottleneck becomes more acute.

Even the most sophisticated artificial intelligence methods available today are still far from being able to reliably review complex artifacts such as some of the submissions in these platforms (or, say, the papers submitted to academic venues). Yet, computational ways can aid human experts in the review process. A complete solution will therefore include computational means; and a better process of dividing the labor between experts, semi-experts, and non-experts. In this paper we demonstrated an open computational approach, which leverages multiple available data from the submission text, as well as from traces of human activity relating to the submission. We borrowed different analytical frameworks, mainly from sociolinguistics, and demonstrated how they can be used to guide the development of predictive models for the task at hand. Importantly, while our models proved to be powerful in the real-life setting of the Climate CoLab, focusing on the specific variables that ended up being salient in our models will miss the message of this paper. Our goal here is to suggest an approach, and report encouraging empirical results. We intend to continue working on modeling proposal success in the Climate CoLab, and to refine this model by observing additional years, and by taking additional variables into account. Further modeling work in different settings will help strengthen the external validity of our results, and provide insight regarding which variables, or families of variables, are important to look at in any settings, and which are context specific. We hope our work will encourage others to join us in pushing open innovation forward.

Acknowledgements

We are grateful to the our mentors and colleagues at the MIT Center for Collective Intelligence and at the Climate CoLab – Tom Malone, Gary Olson, Jeff Nickerson, Rob Laubacher, Laur Fisher, and Mark Klein – and to Abraham Bernstein and James Pennebaker, for their invaluable support and advice. We also thank Klemens Mang, who provided great research assistance. Partial funding for this research was provided by The Swiss National Science Foundation under contract number 200021-143411/1, and sponsors of CCI and the Climate CoLab including the U.S. Army Research Laboratory's Army Research Office (ARO). A.C.B Garcia acknowledges support from CAPES Brazil.

References

Bao, J., Sakamoto, Y., and Nickerson, J. 2011. "Evaluating Design Solutions Using Crowds," Seventeenth Americas Conference on Information Systems, August 4th-7th.

Bennett, J., and Lanning, S. 2007. "The Netflix Prize," KDD Cup and Workshop at the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: Citeseer.

Blohm, I., Riedl, C., Leimeister, J. M., and Krcmar, H. 2011. "Idea Evaluation Mechanisms for Collective Intelligence in Open Innovation Communities: Do Traders Outperform Raters," Proceedings of 32nd International Conference on Information Systems, pp. 1-24.

Bothos, E., Apostolou, D., and Mentzas, G. 2012. "Collective Intelligence with Web-Based Information Aggregation Markets: The Role of Market Facilitation in Idea Management," Expert Systems with Applications (39:1), pp. 1333-1345.

Boudreau, K. J., Guinan, E., Lakhani, K. R., and Riedl, C. 2012. "The Novelty Paradox & Bias for Normal Science: Evidence from Randomized Medical Grant Proposal Evaluations," in: Harvard Business School Technology & Operations Mgt. Unit Working Papers. SSRN.

Boudreau, K. J., Lacetera, N., and Lakhani, K. R. 2011. "Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis," Management Science (57:5), p. 843.

Boudreau, K. J., and Lakhani, K. R. 2013. "Using the Crowd as an Innovation Partner," Harvard Business Review (91:4), pp. 60-69.

Chesbrough, H. W. 2003. "Open Innovation: The New Imperative for Creating and Profiting from Technology (Hardcover)," in Harvard Business School Press Books.

Coleman, M., and Liau, T. L. 1975. "A Computer Readability Formula Designed for Machine Scoring," Journal of Applied Psychology (60:2), p. 283.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popovi, Z., and Foldit_Players. 2010. "Predicting Protein Structures with a Multiplayer Online Game," Nature (466:7307), pp. 756–760.

Dean, D. L., Hender, J. M., Rodgers, T. L., and Santanen, E. L. 2006. "Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation," Journal of the Association for Information Systems (7:10), pp. 646-698.

DuBay, W. H. 2007. Smart Language: Readers, Readability, and the Grading of Text. ERIC.

Gonzales, A. L., Hancock, J. T., and Pennebaker, J. W. 2009. "Language Style Matching as a Predictor of Social Dynamics in Small Groups," Communication Research).

Haller, J. B. A., Bullinger, A. C., and Möslein, K. M. 2011. "Innovation Contests," Business & Information Systems Engineering (3:2), pp. 103-106.

Introne, J., Laubacher, R. J., Olson, G. M., and Malone, T. W. 2011. "The Climate Colab: Large Scale Model-Based Collaborative Planning," Conference on Collaboration Technologies and Systems (CST 2011), Philadelphia, PA: IEEE, pp. 40-47.

Ireland, M. E., and Pennebaker, J. W. 2010. "Language Style Matching in Writing: Synchrony in Essays, Correspondence, and Poetry," Journal of personality and social psychology (99:3), pp. 549-571.

Jeppesen, L. B., and Lakhani, K. R. 2010. "Marginality and Problem-Solving Effectiveness in Broadcast Search," Organization Science (21:5), pp. 1016-1033.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. 1975. "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel," DTIC Document.

Klare, G. R. 1974. "Assessing Readability," Reading Research Quarterly (10:1), pp. 62-102.

Klein, M., and Garcia, A. C. B. 2013. "High-Speed Idea Filtering with the Bag of Lemons," in: SSRN.

Kohavi, R. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," International Joint Conference on Artificial Intelligence, pp. 1137-1145.

Lakhani, K. R., Jeppesen, L. B., Lohse, P. A., and Panetta, J. A. 2007. "The Value of Openness in Scientific Problem Solving," in: Harvard Business School Working Papers.

Malone, T. W., Laubacher, R. J., and Dellarocas, C. 2009. "Harnessing Crowds: Mapping the Genome of Collective Intelligence," in: MIT Sloan Working Papers.

Markoff, J. 2013. "New Test for Computers: Grading Essays at College Level," in: The New York Times.

Marshall, J., and magazie, N. 2012. "Online Gamers Achieve First Crowd-Sourced Redesign of Protein," in: Scientific American.

Morgan, J., and Wang, R. 2010. "Tournaments for Ideas," California management review (52:2), pp. 77-97.

Nagar, Y. 2013. "Designing a Collective-Intelligence System for Evaluating Complex, Crowd-Generated Intellectual Artifacts," 2013 ACM Conference on Computer Supported Collaborative Work (CSCW 2013), San-Antonio, Texas, USA: ACM, pp. 73-76.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. 2011. "Scikit-Learn: Machine Learning in Python," Journal of Machine Learning Research (12), pp. 2825-2830.

Pennebaker, J. W. 2011. "The Secret Life of Pronouns: How Our Words Reflect Who We Are." New York, NY: Bloomsbury Press.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. 2007. "The Development and Psychometric Properties of Liwc2007." Austin, TX: LIWC. Net.

Riedl, C., Blohm, I., Leimeister, J. M., and Krcmar, H. 2013. "The Effect of Rating Scales on Decision Quality and User Attitudes in Online Innovation Communities," International Journal of Electronic Commerce (17:3), pp. 7-36.

Salganik, M. J., and Levy, K. E. C. 2012. "Wiki Surveys: Open and Quantifiable Social Data Collection." Arxiv.

Salminen, J., and Harmaakorpi, V. 2012. "Collective Intelligence and Practice-Based Innovation: An Idea Evaluation Method Based on Collective Intelligence," in Practice-Based Innovation: Insights, Applications and Policy Implications. Springer, pp. 213-232.

Shermis, M. D., and Hamner, B. 2013. "Contrasting State-of-the-Art Automated Scoring of Essays," in Handbook of Automated Essay Evaluation: Current Applications and New Directions, M.D. Shermis and J.C. Burstein (eds.). Routledge, p. 313.

Soukhoroukova, A., Spann, M., and Skiera, B. 2012. "Sourcing, Filtering, and Evaluating New Product Ideas: An Empirical Exploration of the Performance of Idea Markets," Journal of Product Innovation Management (29:1), pp. 100-112.

Swets, J. A. 1988. "Measuring the Accuracy of Diagnostic Systems," Science (240:4857), pp. 1285-1293.

Swets, J. A., and Pickett, R. M. 1982. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. Academic Press.

Terwiesch, C., and Xu, Y. 2008. "Innovation Contests, Open Innovation, and Multiagent Problem Solving," Management Science (54:9), pp. 1529-1543.

Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M. 2008. "Recaptcha: Human-Based Character Recognition Via Web Security Measures," Science (321:5895), p. 1465.

Walter, T. P., and Back, A. 2013. "A Text Mining Approach to Evaluate Submissions to Crowdsourcing Contests," 46th Hawaii International Conference onSystem Sciences (HICSS): IEEE, pp. 3109-3118.

Westerski, A., Dalamagas, T., and Iglesias, C. A. 2013. "Classifying and Comparing Community Innovation in Idea Management Systems," Decision Support Systems (54:3), pp. 1316-1326.

Zweig, M. H., and Campbell, G. 1993. "Receiver-Operating Characteristic (Roc) Plots: A Fundamental Evaluation Tool in Clinical Medicine," Clinical Chemistry (39:4), pp. 561-577.